

Rounding Effects in Record Statistics

G. Wergen,¹ D. Volovik,² S. Redner,² and J. Krug¹

¹*Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany*

²*Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*

(Dated: March 8, 2013)

We analyze record-breaking events in time series of continuous random variables that are subsequently discretized by rounding to integer multiples of a discretization scale $\Delta > 0$. Rounding leads to ties of an existing record, thereby reducing the number of new records. For an infinite number of random variables that are drawn from distributions with a finite upper limit, the number of discrete records is finite, while for distributions with a thinner than exponential upper tail, fewer discrete records arise compared to continuous variables. In the latter case the record sequence becomes highly regular at long times.

PACS numbers: 05.45.Tp, 05.40.-a, 06.20.Dk, 02.50.-r

The statistics of record-breaking events have been widely studied in many contexts, including sports [1], evolutionary biology [2], the theory of spin glasses [3], and the possible role of global warming in the occurrence of record-breaking temperatures [4–9]. Records are defined as the entries in a time series of measurements that exceed all previous values. While the record statistics of independent, identically distributed (iid) random variables (RVs) that are drawn from continuous distributions are well understood [10, 11], the understanding of records drawn from time-dependent distributions [12–14] and from series of correlated RVs [15, 16] is still developing.

Here we address *discreteness effects* on record statistics. Conventionally, records are recorded from variables that are drawn from a continuous distribution. However, in all practical applications, technical limitations cause observations to be discrete, even if the underlying distribution is continuous. In sports or meteorology, distance, time, temperature, or precipitation measurements are always rounded to a certain accuracy [1, 6, 7], resulting in an effective discrete distribution of RVs. Thus ties of existing records can arise, which alters the probability for a record to occur in any given observation (Fig. 1).

For RVs that are explicitly drawn from discrete distributions, the effect of ties strongly affects the number of records [17–21]. For related δ -records and geometric

records, where a new record arises only if the current observation exceeds the current record by a fixed constant δ [21, 22] or by a fixed fraction [23], intriguing statistical properties of records were found for the three universality classes of extreme value statistics (EVS) [24]. However, the consequences of measuring *rounded* record values that are drawn from continuous underlying distributions appears not to have been studied previously.

We consider a set of RVs X_1, \dots, X_N and focus on the probability $P_n \equiv \text{Prob}(X_n > X_1, \dots, X_{n-1})$ that the n^{th} variable in this series is a record. We denote P_n as the *record rate* and $R_n = \sum_{k=1}^n P_k$ as the *record number*. For continuous iid RVs, the universal result is $P_n = \frac{1}{n}$ (see, e.g., [10, 11]). Thus for $n \gg 1$, $R_n \approx \ln n + \gamma$, with $\gamma \approx 0.577\dots$ the Euler constant. We assume that the RVs X_i are discretized in units of a minimal scale Δ . That is, each X_i gets rounded to a value of $X_i^\Delta = k\Delta$. We may consider (i) *rounding down*, with $k = \lfloor X_i/\Delta \rfloor$ and $\lfloor X \rfloor$ the floor function, which gives the largest integer smaller than X , or (ii) *rounding to the nearest lattice point*, with $k = \lfloor X_i/\Delta + \Delta/2 \rfloor$. Because asymptotic results do not depend on the rounding protocol, we will discuss only rounding down. We define the *strong* record rate

$$P_n^\Delta \equiv \text{Prob}(X_n^\Delta > X_1^\Delta, \dots, X_{n-1}^\Delta), \quad (1)$$

in which ties caused by the discretization are *not counted* as new records. Thus not only X_n , but also the rounded value X_n^Δ has to be larger than all previous RVs for a new record to occur (Fig. 1).

General theory, asymptotic results. For iid RVs X_i drawn from a distribution with probability density $f(x)$ and cumulative distribution $F(x) = \int^x dy f(y)$, the record rate is obtained from $P_n = \int dx f(x) F^{n-1}(x)$ [11]. For any continuous density $f(x)$, this integral gives the universal behavior mentioned above, $P_n = \frac{1}{n}$. However, if the measurement X_i is rounded down to X_i^Δ , the inte-

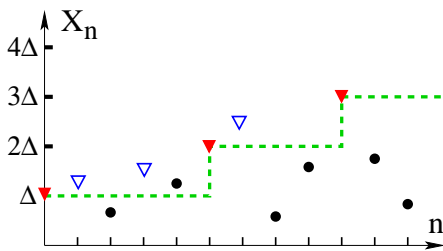


FIG. 1: (color online) Effect of rounding down records with discretization unit Δ . Inverted triangles indicate records, with those that survive after rounding shown solid. The dashed line shows the evolution of the rounded record value.

gral for P_n breaks into the sum

$$\begin{aligned} P_n^\Delta &= \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx f(x) \right] F^{n-1}(k\Delta), \\ &= \sum_k [F((k+1)\Delta) - F(k\Delta)] F^{n-1}(k\Delta). \end{aligned} \quad (2)$$

This gives the strong record rate from continuous RVs that are rounded down to the closest integer multiple of Δ . We emphasize that in the practically more relevant case where record values are rounded either up or down to the closest integer multiple of Δ , the record rate has the same statistical properties as those from only rounding down. We now give asymptotic results for P_n^Δ for the three basic classes of EVS [24]: Weibull (distributions with a finite upper limit), Gumbel (unbounded upper tail decaying faster than any power law), and Fréchet (power-law upper tail). Our asymptotic approximations for the discrete record rate P_n^Δ for these classes of EVS agree well with numerical results.

Weibull class: For illustration, we start with the uniform distribution: $f(x) = 1$ for $x \in [0, 1]$ and 0 otherwise. For discretization scale $\Delta = \frac{1}{L}$, with integer-valued $L > 1$, Eq. (2) reduces to:

$$P_n^\Delta = \sum_{k=1}^{\frac{1}{\Delta}-1} \Delta (k\Delta)^{n-1} = \Delta^n H_{\frac{1}{\Delta}-1, n-1}, \quad (3)$$

where $H_{m,n}$ is the m^{th} harmonic number of power n . At some point in the time series of RVs, a record with a rounded value $1 - \Delta$ occurs; this is necessarily the *last record*. For a fine discretization scale, $\Delta \ll 1$, the sum in (3) can be replaced by an integral to give $P_n^\Delta \approx \frac{1}{n} (1 - \Delta)^n$. Thus for any $\Delta > 0$, P_n^Δ no longer decays as $\frac{1}{n}$, but instead approaches zero exponentially with n — rounding strongly depresses the asymptotic record rate for the uniform distribution.

A more general example of the Weibull EVS class is $f(x) = \xi(1-x)^{\xi-1}$, with $\xi > 0$ and $x \in [0, 1]$. By expanding Eq. (2) to second order for $\Delta \ll 1$, we find

$$\begin{aligned} P_n^\Delta &\approx \int_1^{\frac{1}{\Delta}-1} dk \left[(1-k\Delta)^\xi - (1-(k+1)\Delta)^\xi \right] \\ &\quad \times [1 - (1-k\Delta)^\xi]^{n-1}, \\ &\approx \begin{cases} \frac{1}{n} \left[1 - n\Delta^\xi - \frac{\Delta^\xi}{2} \Gamma(2 - \frac{1}{\xi}) n^{1/\xi} \right], & n\Delta^\xi \ll 1, \\ \frac{1}{n} \exp(-n\Delta^\xi), & n\Delta^\xi \gg 1. \end{cases} \end{aligned} \quad (4)$$

Since the underlying distribution has a bounded support, the total number of records is again finite. The results in (4) reproduce those found for the uniform distribution.

Gumbel class: As a basic example, we treat the exponential distribution $f(x) = e^{-x}$. For $n \gg 1$ we replace

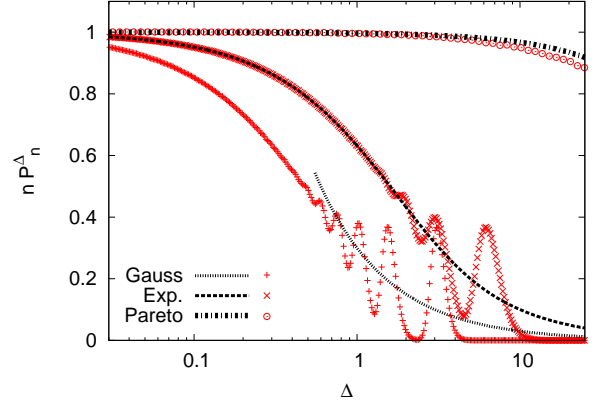


FIG. 2: (color online) Scaled record rate nP_n^Δ for $n = 1000$ for the Gaussian, exponential, and Pareto (with $\mu = 1.2$) distributions. Without rounding, $P_n = \frac{1}{n}$. Simulations (symbols) are averaged over 10^6 time series and over $975 \leq n \leq 1025$ to smooth the data. Analytical predictions (curves) are shown for comparison. For the origin of the peaks for the Gaussian and exponential distributions, see the text following Eq. (14).

the sum in Eq. (2) by an integral and find

$$P_n^\Delta \approx \sum_{k=1}^{\infty} e^{-k\Delta} (1 - e^{-k\Delta})^n \approx \frac{1}{n\Delta} (1 - e^{-\Delta}) \quad (5)$$

for arbitrary $\Delta \geq 0$, in agreement with findings for the geometric distribution in Ref. [18] and with our simulations (Fig. 2). For $\Delta \ll 1$, (5) reduces to $P_n^\Delta \approx \frac{1}{n} (1 - \frac{\Delta}{2})$, while for $\Delta \gg 1$, $P_n^\Delta \approx \frac{1}{n\Delta}$. In contrast to the Weibull class, P_n^Δ asymptotically decays as $\frac{1}{n}$ for arbitrary Δ .

For the Gaussian distribution $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, with unit standard deviation, we find that as $n \rightarrow \infty$

$$\begin{aligned} P_n^\Delta &\approx \frac{1}{2} \int dx \left[\operatorname{erfc} \left(\frac{k\Delta}{\sqrt{2}} \right) - \operatorname{erfc} \left(\frac{(k+1)\Delta}{\sqrt{2}} \right) \right] F(x)^{n-1}, \\ &\approx \frac{1}{\Delta} \int dx \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2} F(x)^{n-1}. \end{aligned} \quad (6)$$

For $n \rightarrow \infty$ we evaluate this integral by the Laplace method by expanding the integrand about $x^* = \ln(n^2/2\pi)$, where x^* is the mean value of the n^{th} record. After some calculation, we obtain

$$P_n^\Delta \approx \frac{1}{n\Delta} \left[\sqrt{\ln \left(\frac{n^2}{2\pi} \right)} \right]^{-1}. \quad (7)$$

Thus the record rate decays slightly faster than $\frac{1}{n}$ (Fig. 2). Correspondingly, $R_n^\Delta \propto \Delta^{-1} (\ln n)^{1/2}$, which diverges weakly as $n \rightarrow \infty$.

Fréchet class: A representative for this class is the Pareto distribution $f(x) = \mu x^{-\mu-1}$, with $x > 1$ and $\mu > 0$. Using again Eq. (2), the asymptotic record rate P_n^Δ is

$$P_n^\Delta \approx \frac{1}{n} \left[1 - \frac{\Delta}{2} \mu \Gamma \left(2 + \frac{1}{\mu} \right) n^{-1/\mu} \right]. \quad (8)$$

In contrast to the two previous classes, the effect of the rounding is negligible, as $P_n^\Delta \rightarrow P_n$ for $n \rightarrow \infty$ and arbitrary Δ (Fig. 2).

Small- Δ regime. We now focus on the effects of rounding when the discretization scale is small ($\Delta \ll 1$) for fixed n . Here we find a useful analogy between the effect of a linear drift in RVs [13] and the effect of rounding, and we adapt methods developed for the former problem to help elucidate rounding effects. For small Δ the general expression (2) for P_n^Δ simplifies to

$$\begin{aligned} P_n^\Delta &= \sum_k \left[\int_{k\Delta}^{(k+1)\Delta} dx f(x) \right] F^{n-1}(k\Delta), \\ &= \int dx f(x) F^{n-1}([x]_\Delta), \\ &\approx \frac{1}{n} - n \int dx (x - [x]_\Delta) f^2(x) F^{n-2}(x). \end{aligned} \quad (9)$$

Here $[x]_\Delta$ is defined as the largest integer multiple of Δ that is smaller than x . Thus, in the second line, $k\Delta = [x]_\Delta$ for $k\Delta \leq x < (k+1)\Delta$, which obviates writing the sum. In the last step, we expand to first order in the quantity $x - [x]_\Delta$ and employ the crude assumption that, on average, $x - [x]_\Delta \approx \frac{\Delta}{2}$ to give

$$P_n^\Delta \approx \frac{1}{n} \left(1 - \frac{\Delta}{2} n^2 \mathcal{I}_n \right), \quad (10)$$

where $\mathcal{I}_n \equiv \int dx f^2(x) F^{n-2}(x)$. The approximation underlying (10) is valid if $n^2 \Delta \mathcal{I}_n \ll 1$. The quantity \mathcal{I}_n appears in record statistics that arise from continuous RVs with a linear drift [13], whose behavior is known for a wide range of distributions. In the following we use the results from [13] to determine P_n^Δ in the small- Δ regime.

Weibull and Fréchet classes: For the distribution $f(x) = \xi(1-x)^{\xi-1}$ introduced above, the approximation given by Eq. (10) is useful for $\xi > 1$ and we find, for $n\Delta^\xi \ll 1$,

$$P_n^\Delta \approx \frac{1}{n} \left[1 - \frac{\Delta^\xi}{2} \Gamma\left(2 - \frac{1}{\xi}\right) n^{1/\xi} \right], \quad (11)$$

which, for $n\Delta^\xi \ll 1$ and $\xi > 1$, agrees with the result derived from our general approach in Eq. (4). Similarly, for the Pareto distribution we recover Eq. (8).

Gumbel class: For the exponential distribution, we find $P_n^\Delta \approx \frac{1}{n} \left(1 - \frac{\Delta}{2} \right)$, which agrees with the small- Δ behavior of Eq. (5). For the Gaussian distribution, the small- Δ approximation allows us to obtain a new expression for the record rate when $\sqrt{\ln n} \ll \Delta^{-1}$,

$$P_n^\Delta \approx \frac{1}{n} \left[1 - \frac{2\Delta\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{n^2}{8\pi}\right)} \right]. \quad (12)$$

The regime $\sqrt{\ln n} \ll \Delta^{-1}$ is not accessible through the general approach and this range is particularly important for applications, such as in climatology [7]. For $n \gg 1$

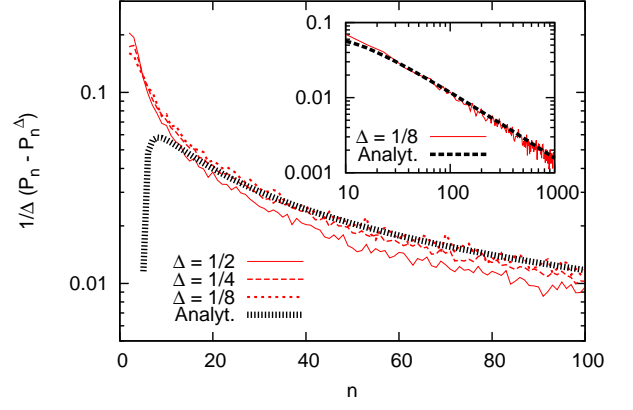


FIG. 3: (color online) Simulations of P_n^Δ for Gaussian RVs in the regime $\sqrt{\ln n} \ll \frac{1}{\Delta}$. Thin curves are $\frac{1}{\Delta} (P_n - P_n^\Delta)$ for $\Delta = \frac{1}{2}, \frac{1}{4}$ and $\frac{1}{8}$ and $n \in [0, 100]$. For each Δ , 10^6 time series were simulated. The thick dashed curve depicts the analytical prediction Eq. (15). Inset shows the same analysis for $\Delta = \frac{1}{8}$ with $n \in [1, 1000]$.

and $\Delta \ll 1$, Eq. (12) reproduces the numerical simulation values for P_n^Δ very accurately (Fig. 3).

Large- Δ regime. For Gumbel-class distributions that decay at least exponentially fast near the upper limit, we can provide an alternative description for the record number R_n^Δ . For these distributions, it is known that the average spacings between the record events do not increase in time for large n [11]. Therefore, we may choose a sufficiently large value of Δ that almost all records are suppressed because of ties. It then follows that all discrete values $k\Delta$ (with $k \geq 0$) will eventually be record values and R_n^Δ is just the sum over the probabilities that a record has already occurred for a certain value $k\Delta$. The corresponding probabilities $\Pi_n(k)$ for record value $k\Delta$ are given by $\Pi_n(k) \approx 1 - F(k\Delta)^{n-1}$, which leads to

$$R_n^\Delta \approx \sum_{k=0}^{\infty} \Pi_n(k) \approx 1 + \sum_{k=1}^{\infty} [1 - F(k\Delta)^{n-1}]. \quad (13)$$

For elementary Gumbel distributions, interesting properties emerge from $\Pi_n(k)$. For a small n and large $k\Delta$, it is obvious that $\Pi_n(k) \approx 0$. Conversely, for large n and arbitrary $k\Delta$ eventually $\Pi_n(k) \approx 1$, since $F(k\Delta) < 1$ for finite $k\Delta$.

We now estimate the regime where $\Pi_n(k)$ switches between 0 and 1; this condition also determines the point where the mean record number switches from $k-1$ to k . Since $\Pi_n(k)$ will never be exactly 0 or 1, we seek the time n , where $\Pi_n(k)$ is either smaller than ϵ ($n=n_-$) or larger than $1-\epsilon$ ($n=n_+$) for small $\epsilon \ll 1$. By elementary means we find

$$n_- < \frac{\ln \epsilon}{\ln [F(k\Delta)]}, \quad n_+ > \frac{\epsilon}{-\ln [F(k\Delta)]}. \quad (14)$$

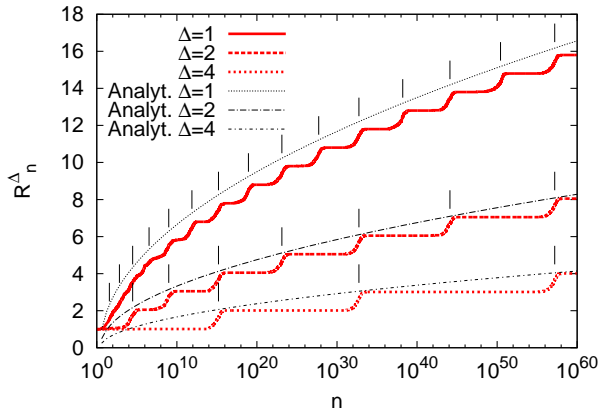


FIG. 4: (color online) Record number R_n^Δ for Gaussian RVs for $\Delta = 1, 2, 4$. Data (bold lines) are based on 100 realizations with a maximal $n = 10^{60}$. For $n > 10^6$ we used an algorithm that directly simulates record events by sampling both the distribution and the waiting time of the $(k+1)^{\text{st}}$ record from the value of the k^{th} record. Thin lines show the asymptotic behavior predicted by Eq. (15). The vertical lines show the steps predicted by $n \approx \sqrt{2\pi k \Delta} e^{(k\Delta)^2/2}$.

Evidently $\Pi_n(k)$ switches between 0 and 1 when n is between n_- and n_+ , where n_- and n_+ are both proportional to $[\ln(F(k\Delta))]^{-1}$. For the exponential distribution, for example, we find that $n_- = \epsilon e^{k\Delta}$ and $n_+ = \ln(1/\epsilon) e^{k\Delta}$, so the k^{th} record will occur at a time proportional to $e^{k\Delta}$, leading to a mean record number of $R_n^\Delta \approx \frac{1}{\Delta} \ln n$. In the large $k\Delta$ regime, records occur in an ordered fashion and are well separated from each other. The $(k+1)^{\text{st}}$ record occurs at time $e^{(k+1)\Delta}$, which for $\Delta \gg 1$, is much later than the time of the k^{th} record. Thus the mean record number undergoes a step-like periodicity when plotted against e^n . For the Gaussian distribution, the same approach now predicts that $\Pi_n(k)$ switches for $n \approx \sqrt{2\pi k \Delta} e^{k^2 \Delta^2/2}$ (Fig. 4). For large $k\Delta$ and large n , the mean record number becomes

$$R_n^\Delta \approx \sum_{k=0} \Pi_n(k) \approx \frac{1}{\Delta} \sqrt{\ln \left(\frac{n^2}{2\pi} \right)}, \quad (15)$$

which was already obtained with the general approach above and confirms the validity of the form for R_n^Δ given in Eq. (13). The step periodicity in R_n^Δ is the source of the observed peaks (Fig. 2) in the record rate P_n^Δ as a function of Δ for exponential and Gaussian distributions.

Conclusions. We determined how rounding down continuous random variables affects the statistics of records. Our results directly apply to the practical situation where continuous variables are rounded either up or down to the closest integer multiple of a fixed discretization scale Δ .

For distributions with bounded support, rounding leads to an exponential decay of the record rate, P_n^Δ , and an asymptotically finite record number. In contrast, for

power-law distributions, the effect of rounding becomes negligible for $n \rightarrow \infty$ and $P_n^\Delta \rightarrow \frac{1}{n}$ independent of Δ . In the intermediate Gumbel class, the behavior is more subtle. For the exponential distribution, P_n^Δ decays as $\frac{1}{n}$ with a Δ -dependent prefactor, while for the general distribution $f(x) \propto \exp(-|x|^\beta)$ with $\beta > 1$, the record rate decays as $n^{-1} \ln(n)^{1/\beta-1}$.

For underlying distributions that decay at least exponentially, the record sequence becomes ordered at long times, in marked contrast to independent record events from continuous iid RVs [10, 11]. While correlations between record events have been previously observed for RVs that are drawn from drifting [14] or broadening [12] distributions, the effect of rounding is much stronger and renders record events predictable on a time scale that grows exponentially (or faster) with record number.

To illustrate that rounding effects have an observationally significant influence on records, we analyzed 50 years of daily temperatures from 361 U.S. weather stations [25] along the lines of [7]. The measurements were reported in integer units of $\Delta = 1^\circ\text{F}$ and we considered all 361×365 time series for the individual calendar days with an average standard deviation of $\sigma \approx 8.9^\circ\text{F}$. Only 75% of the weak upper (ties allowed) and 78% of the weak lower records were also strong records (no ties), in good agreement with the value of 79% predicted by our analytical result in Eq. (12). In this example the effect of ties on the record rate has a similar magnitude as that of the small warming trend in the data (cf. [5–7]). Thus rounding effects should be carefully accounted for if one wishes to use record statistics to detect secular trends in data, such as global warming.

GW acknowledges financial support from Friedrich-Ebert-Stiftung and BCGS as well as the kind hospitality of the Center for Polymer Studies in the early stages of this work. DV and SR thank NSF grant DMR-0906504 for partial financial support of this research. We thank O. Pulkkinen for making us aware of Refs. [17, 18].

-
- [1] D. Gembris, J. G. Taylor, and D. Suter, *Nature* **417**, 506 (2002); D. Gembris, J. G. Taylor, and D. Suter, *J. Appl. Stat.* **34**, 529 (2007).
 - [2] J. Krug and K. Jain, *Physica A* **358**, 1 (2005).
 - [3] L. P. Oliveira et al., *Phys. Rev. B* **71**, 104526 (2005); P. Sibani, G. F. Rodriguez, and G. G. Kenning, *Phys. Rev. B* **74**, 224407 (2006).
 - [4] G. W. Bassett, *Climatic Change* **21**, 303 (1992); R. E. Benestad, *Climate Research* **25**, 3 (2003).
 - [5] S. Redner and M. R. Petersen, *Phys. Rev. E* **74**, 061114 (2006).
 - [6] G. A. Meehl et al., *Geophys. Res. Lett.* **36**, L23701 (2009).
 - [7] G. Wergen and J. Krug, *Europhys. Lett.* **92**, 30008 (2010).
 - [8] W. I. Newman, B. D. Malamud, and D. L. Turcotte,

- Phys. Rev. E **82**, 066111 (2010).
- [9] S. Rahmstorf and D. Coumou, Proc. Natl. Acad. Sci. USA **108**, 17905 (2011).
 - [10] N. Glick, Amer. Math. Monthly **85**, 2 (1978).
 - [11] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, Records. Wiley, New York (1998); V. B. Nevzorov, Theor. Probab. Appl. **32**, 201 (1987).
 - [12] J. Krug, J. Stat. Mech. P07001 (2007),
 - [13] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech. P10013 (2010).
 - [14] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011); J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).
 - [15] S. N. Majumdar and R.M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008); S. Sabhapandit, Europhys. Lett. **94**, 20003 (2011).
 - [16] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. E **83**, 051109 (2011).
 - [17] W. Vervaat, Stochastic Processes Appl. **1**, 317 (1973).
 - [18] H. Prodinger, Discrete Mathematics **153**, 253 (1996).
 - [19] R. Gouet, F. J. López, and G. Sanz, Adv. Appl. Prob. **37**, 118 (2005).
 - [20] E. S. Key, J. Theor. Probab. **18**, 99 (2005).
 - [21] R. Gouet, F. J. López, and G. Sanz, Bernoulli **13**, 754 (2007).
 - [22] N. Balakrishnan, K. Balasubramanian, and S. Panchapakesan, J. Appl. Statist. Sci. **4**, 123 (1996).
 - [23] R. Gouet, F. J. López, and G. Sanz, J. Stat. Mech. P01005 (2012); I. Eliazar, Physica A **348**, 181 (2005).
 - [24] E. J. Gumbel, National Bureau of Standards Applied Mathematics Series **33** (1954); L. de Haan and A. Ferreira, *Extreme Value Theory - An Introduction*, (Springer, New York, 2006).
 - [25] M. J. Menne, C. N. Williams Jr., and R. S. Vose, National Climatic Data Center, National Oceanic and Atmospheric Administration (2010).